Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

How well can carcinogenicity be predicted by high throughput "characteristics of carcinogens" mechanistic data?

Richard A. Becker [a, *], David A. Dreier [b], Mary K. Manibusan [c], Louis A. (Tony) Cox [d], Ted W. Simon [e], James S. Bus [f]

[a] *American Chemistry Council, 700 Second St., NE, Washington DC 20002, USA*
[b] *Center for Environmental & Human Toxicology, University of Florida, Gainesville, FL, USA*
[c] *Exponent, Washington DC, USA*
[d] *Cox Associates, Denver, CO, USA*
[e] *Ted Simon LLC, Winston, GA, USA*
[f] *Exponent, Alexandria, VA, USA*

ARTICLE INFO

Article history:
Received 28 February 2017
Received in revised form
28 August 2017
Accepted 29 August 2017
Available online 1 September 2017

Keywords:
Cancer hazard classification
IARC
Cancer prediction modeling
High throughput screening (HTS)

ABSTRACT

IARC has begun using ToxCast/Tox21 data in efforts to represent key characteristics of carcinogens to organize and weigh mechanistic evidence in cancer hazard determinations and this implicit inference approach also is being considered by USEPA. To determine how well ToxCast/Tox21 data can explicitly predict cancer hazard, this approach was evaluated with statistical analyses and machine learning prediction algorithms. Substances USEPA previously classified as having cancer hazard potential were designated as positives and substances not posing a carcinogenic hazard were designated as negatives. Then ToxCast/Tox21 data were analyzed both with and without adjusting for the cytotoxicity burst effect commonly observed in such assays. Using the same assignments as IARC of ToxCast/Tox21 assays to the seven key characteristics of carcinogens, the ability to predict cancer hazard for each key characteristic, alone or in combination, was found to be no better than chance. Hence, we have little scientific confidence in IARC's inference models derived from current ToxCast/Tox21 assays for key characteristics to predict cancer. This finding supports the need for a more rigorous mode-of-action pathway-based framework to organize, evaluate, and integrate mechanistic evidence with animal toxicity, epidemiological investigations, and knowledge of exposure and dosimetry to evaluate potential carcinogenic hazards and risks to humans.

© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Cancer pathogenesis includes a number of molecular and biological hallmarks (Hanahan and Weinberg, 2000, 2011). This recognition has contributed to an increased emphasis on integrating mechanistic data and knowledge of mode of action (MOA) into judgments of whether a chemical poses a cancer risk to humans. For many chemicals, the scientific basis for assessing human carcinogenic hazard, the selection of a dose-response extrapolation method, and the quantification of cancer risks at environmental levels of exposures all depend on evaluation and interpretation of mechanistic evidence.

Notably, the choice of the likely operative mode of action, which dictates the low-dose extrapolation method, has significant consequences for risk assessment and chemical regulation. For example, USEPA's *Guidelines for Carcinogen Risk Assessment* (USEPA, 2005) invokes a linear extrapolation approach as a default "in the absence of sufficiently, scientifically justifiable mode of action information." California's Proposition 65 regulatory provisions routinely use determinations of cancer hazard by the International Agency for Research on Cancer (IARC) that rely on mechanistic information. In California, should IARC determine there is sufficient evidence of carcinogenicity in either humans or laboratory animals, the chemical is listed as "known to the state to cause cancer" with accompanying warnings to the public and regulatory limits corresponding to a linear low-dose extrapolation intake level posing a 10^{-5} lifetime risk of cancer (https://oehha.ca.gov/proposition-65/general-info/proposition-65-plain-language). For many chemicals,

[*] Corresponding author.
E-mail address: rick_becker@americanchemistry.com (R.A. Becker).

http://dx.doi.org/10.1016/j.yrtph.2017.08.021

0273-2300/© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

no consensus exists regarding integration of the available mechanistic evidence; hence, the positions of organizations other than IARC often differ from those of IARC and from each other about whether the mechanistic data are sufficient to support a determination of human cancer hazard (Dourson et al., 2013).

USEPA's *Guidelines for Carcinogen Risk Assessment* (USEPA, 2005) emphasize the importance of mode of action (MOA) analysis in understanding human cancer risk and this document provides a framework for the analysis of mechanistic data. Indeed, this framework is consistent with the approaches of the World Health Organization and the International Life Sciences Institute Risk Sciences Institute (Meek et al., 2013, 2014; Boobis et al., 2006; Sonich-Mullin et al., 2001). The framework provides a rigorous and structured approach for determining human relevance by evaluating evidence in both humans and animals regarding causally linked key events in pathways leading to carcinogenic effects; in other words, mode of action. To improve scientific justification for the use of MOA in hazard characterization and dose-response analysis, Becker et al. (2017) have developed a simple scoring method for assessing confidence in the supporting mechanistic data for hypothesized mode(s) of action.

USEPA's ToxCast™ program and the Tox21 federal agency collaboration have been pioneers in high-throughput *in vitro* screening (HTS), and these programs have created a wealth of mechanistic data. Despite much effort, it is unclear whether HTS data can inform us of potential hazards to humans (Hill et al., 2017; Cox et al., 2016; Kleinstreuer et al., 2013). The use of these HTS data have been incentivized by being publicly available and IARC has recently used the data in cancer hazard evaluations. Unfortunately, IARC does not appear to fully appreciate the difficulty in evaluating and integrating such *in vitro* data for hazard assessment, despite IARC's guidelines that encourage the Working Groups to attempt to identify possible mechanisms from human, animal, and *in vitro* data (IARC, 2016c). Recently, IARC (Guyton, 2015) has indicated how mechanistic data can elevate causal determinations from probable (IARC Group 2A) to known human carcinogen (Group 1), from possible (Group 2B) to probable (Group 2A) and from not classifiable (Group 3) to either possible (Group 2B) or probable (Group 2A). The IARC framework (Guyton, 2015) also indicates that if there is strong evidence that a mechanism in animals does not operate in humans, a substance could be downgraded from possible (Group 2B) to not classifiable (Group 3). It is imperative that such determinations be based on rigorous objective and transparent assessments and integration of mechanistic data.

In this regard, IARC has developed a judgment-based grouping of mechanistic evidence for use in implicit causal inference. This approach relies on ten mechanism-based key characteristics of known carcinogens (Table 1) (Smith et al., 2016). This approach has been illustrated for two example chemicals (i.e., benzene and polychlorinated biphenyls) using chemical-specific datasets (Lauby-Secretan et al., 2013, 2016; McHale et al., 2012). Recently, IARC Working Groups have extended this inference approach to ToxCast/Tox21 assay data by assigning various assays to seven of the ten key characteristics of carcinogens using expert judgment and then incorporating ToxCast/Tox21 results as part of the evaluation of mechanistic evidence (Guyton, 2015; Loomis et al., 2015; IARC, 2016a,b; IARC, 2017). However, IARC has yet to provide any specific details of the Working Groups' assessments of the performance of these mechanistic assays including assay relevance, reproducibility/reliability, specificity and domain of applicability and predictivity. To an outside observer, the IARC process currently appears as an *ad hoc* subjective evaluation of the mechanistic evidence by each individual IARC working group without accompanying *a priori* science-based ground rules or systematic guidance; this evidence includes ToxCast/Tox21 and other data grouped into each of 7/10 key characteristics of carcinogens. The Working Groups then assign a descriptor of "strong," "moderate," or "weak" to the mechanistic evidence, and these descriptors may then be used to alter determinations of potential human cancer hazard.

IARC's predictions of cancer hazard could be significantly improved by adopting a more scientifically robust approach for integrating mechanistic evidence in lieu of the current *ad hoc* method. This improvement would also apply to use of mechanistic data such as ToxCast/Tox21 results in hazard identification and risk assessment programs in other organizations (e.g., USEPA, NTP/NIEHS, EChA, etc.) to strengthen the scientific foundation of decisions that rely on such assessments. Resulting regulatory decisions would in time come to reflect this knowledge of key molecular and cellular responses associated with cancer pathogenesis revealed by these 21st century technologies. The intent of this paper is to evaluate the approach used by IARC Working Groups involving the reliance of ToxCast/Tox21 results associated with various key characteristics of carcinogens to inform cancer hazard determinations.

2. Methods

A schematic summarizing the workflow of this investigation (data acquisition, analysis and interpretation) is presented in Fig. 1. Data rich chemicals previously evaluated by USEPA's Office of Pesticide Programs' Cancer Assessment Review Committee (OPP/CARC) for carcinogenic hazard, largely based on GLP-conducted rodent cancer bioassays using USEPA or OECD test guidelines, were used to test the hypothesis of whether ToxCast/Tox21 mechanistic studies indicating bioactivity in one or more IARC key characteristics can reliably distinguish carcinogens from non-carcinogens—the former being chemicals classified as posing a carcinogenic hazard to humans and the latter being those without a human cancer hazard. Since USEPA guidelines have been updated over the years, different guidelines have been used in classifying the chemicals depending on the date of the assessment. Hence, for the statistical comparisons and prediction analyses, two groups of chemicals were used: those classified as having human cancer hazard potential (i.e., "Known/Likely" or Group B) as positives and substances not posing a carcinogenic hazard (i.e., "Not Likely" or Group E) as negatives.

The rationale for drawing the dataset from USEPA classifications and not IARC classifications is as follows. First and foremost, it is not possible to designate sufficient negatives from IARC because IARC Group 3 substances are described as "Not Classifiable as to Its Carcinogenicity to Humans" and IARC has only designated a single substance as "Group 4, Probably Not Carcinogenic to Humans" (http://monographs.iarc.fr/ENG/Classification/IARC). Therefore, since USEPA is considering using the ten key characteristics of carcinogens within their evolving systematic review procedures (USEPA, 2015; USEPA, 2017), and USEPA has an extensive set of robust, data-rich chemical evaluations within its Office of Pesticide Programs (OPP), the dataset for analysis was drawn from USEPA's Annual Cancer Report 2016 (USEPA, 2016). This report summarizes the weight of evidence classifications of official regulatory determinations of carcinogenic hazards to humans (USEPA, 2016). ToxCast/Tox21 data are available for 194 substances classified by USEPA as "Not Likely to be Carcinogenic to Humans" and "Group E–Evidence of Noncarcinogenicity for Humans." Thus, for this dataset, there are a sufficient number of substances that can be used as "negatives" for prediction modeling and analysis. Only one substance with ToxCast/Tox21 data (diuron) in the USEPA 2016 Annual Cancer Report is characterized as "Known/Likely" and this substance was included in our analysis as a "positive," since it was classified as "Likely." While the "positives" dataset could have been

Table 1
The 10 key characteristics of carcinogens described by Smith et al., (2016) and used in IARC Monographs 112 (IARC, 2017) and 113 (IARC, 2016b).

Key Carcinogen Characteristic	IARC's Assignment and Use of ToxCast Data
1) Is Electrophilic or Can Be Metabolically Activated to Electrophiles	9 ToxCast assays associated with this characteristic[a]
2) Is Genotoxic	31 ToxCast assays associated with this characteristic
3) Alters DNA Repair or Causes Genomic Instability	No ToxCast assays could be associated with this characteristic
4) Induces Epigenetic Alterations	11 ToxCast assays associated with this characteristic
5) Induces Oxidative Stress	18 ToxCast assays associated with this characteristic
6) Induces Chronic Inflammation	45 ToxCast assays associated with this characteristic
7) Is Immunosuppressive	No ToxCast assays could be associated with this characteristic
8) Modulates Receptor-Mediated Effects	92 ToxCast assays associated with this characteristic
9) Causes Immortalization	No ToxCast assays could be associated with this characteristic
10) Alters Cell Proliferation, Cell Death or Nutrient Supply	68 ToxCast assays associated with this characteristic

[a] All 9 assays are p53 constructs. Although these p53 ToxCast data were used in IARC Monograph 112, in the subsequent Monograph 113, the p53 assay results and the characteristic Is Genotoxic were not included in the Monograph analysis section "Aligning in-vitro assays to the 10 "key characteristics" of known human carcinogens."

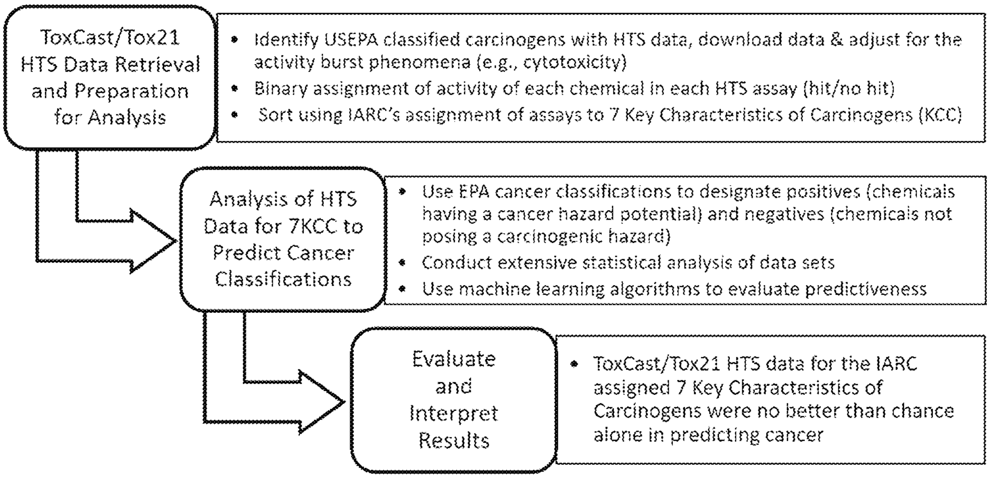

ToxCast/Tox21 HTS Data Retrieval and Preparation for Analysis
Identify USEPA classified carcinogens with HTS data, download data & adjust for the activity burst phenomena (e.g., cytotoxicity)
Binary assignment of activity of each chemical in each HTS assay (hit/no hit)
Sort using IARC's assignment of assays to 7 Key Characteristics of Carcinogens (KCC)
Analysis of HTS Data for 7KCC to Predict Cancer Classifications
Use EPA cancer classifications to designate positives (chemicals having a cancer hazard potential) and negatives (chemicals not posing a carcinogenic hazard)
Conduct extensive statistical analysis of data sets
Use machine learning algorithms to evaluate predictiveness
Evaluate and Interpret Results
ToxCast/Tox21 HTS data for the IARC assigned 7 Key Characteristics of Carcinogens were no better than chance alone in predicting cancer

Fig. 1. Schematic summarizing the workflow of this investigation (data acquisition, analysis and interpretation).

augmented by including the chemicals in IARC Group 1 and Group 2 that are not part of USEPA's Annual Cancer Report 2016, we elected not to do so, primarily because of known differences in the weight of evidence procedures for cancer evaluation used by IARC and USEPA. Most importantly, USEPA uses a transparent and comprehensive weight of evidence process that typically includes public comment and independent peer review procedures, whereas IARC decisions employ strength of evidence evaluation procedures (Boobis et al., 2016) within closed meetings of selected invited experts.

2.1. Assays and key characteristics

We used the IARC Working Group's assignment of specific assays to the key characteristics in this analysis (IARC, 2017). Notably, we did not conduct an independent evaluation of the assays nor did we independently assign assays to the key characteristics. The ten key characteristics (Smith et al., 2016) and the number of assays assigned to each key characteristic (IARC, 2016b; 2017) are summarized in Table 1. IARC (2016c, 2017) identified relevant ToxCast/Tox21 assays for 7 of the 10 IARC key characteristics; no ToxCast/Tox21 assays were associated by IARC with the following characteristics: alters DNA repair or causes genomic instability; is immunosuppressive, and; causes immortalization. Hence in this analysis we only address these seven key characteristics: (1) Is electrophilic or can be metabolically activated to electrophiles; (2) Is genotoxic; (4) Induces epigenetic alterations; (5) Induces oxidative stress; (6) Induces chronic inflammation; (8) Modulates receptor-mediated effects; and (10) Alters cell proliferation, cell death or nutrient supply.

ToxCast/Tox21 data for 54 chemicals meeting the criteria as positives and 194 chemicals meeting the criteria as negatives were compiled from USEPA's online ToxCast/Tox21 summary files and grouped to each key characteristic for the seven key characteristics (data accessed and downloaded on 9 August 2016).

2.2. Accounting for cytotoxicity

Cytotoxicity and non-specific responses present major challenges for interpreting ToxCast/Tox21 chemical-assay combinations. To date, all of the IARC Working Groups' analyses have been conducted without considering the concentration-dependent burst activity phenomenon described by Judson et al. (2016); this phenomenon is observed both in cell-free biochemical assays and as cytotoxicity in cell-based assays. Recently, several cytotoxicity assays within ToxCast have been used to screen individual substances and develop concentration thresholds to distinguish specific activity of a chemical from high concentration, non-specific effects (Judson et al., 2016). IARC did not adjust for cytotoxicity, and we initially analyzed the ToxCast/Tox21 dataset without this adjustment for consistency with IARC (see Supplemental Material).

Based on comments and suggestions on the initial analysis, we conducted and herewith present the identical analysis conducted on a dataset adjusted for the activity burst phenomena. The adjustment selected hits below a cytotoxicity threshold previously defined by Judson et al., (2016). For each chemical, cytotoxicity was defined by 33 related assays, where a hit in 2 or more assays was designated as an active burst response. The burst region for each chemical was equal to 3 times the global median absolute deviation, which also comprised the lower bound cytotoxicity threshold. Hits below this threshold were selected to adjust for burst activity.

Not all assays were run for each chemical, hence different numbers of assays were assigned to each characteristic and the overall dataset necessarily contained null data. Most of the analyses were conducted on the percent activity for each chemical within each key characteristic. This was determined by dividing the hit count by the total number of tested assays. For consistency with some of the analyses of Hill et al. (2017), we also used assay hits per chemical as a measure of overall activity. A summary of these data is provided in Excel spreadsheets, as part of the Supplemental Material.

2.3. Statistical analysis methods

First, we compared the distribution of assay hits per chemical over all the assays for substances classified as having human cancer hazard potential to substances not posing a carcinogenic hazard using the Kolmogorov two-sample test (e.g., Hill et al., 2017). We next attempted to segregate both the percent activity results and hit calls per chemical by key characteristic and explored differences between chemicals based on assay results using plots of empirical cumulative distribution functions (CDFs) for each characteristic. We then examined whether principal component analysis could reveal a difference between positives and negatives based on hit calls for all seven key characteristics. To explore similarities/differences of responses of positives and negatives based on key characteristic or individual assay hit calls, we used the Jaccard similarity index, a method similar to the Tanimoto index for cheminformatic similarity (https://pubchem.ncbi.nlm.nih.gov/score_matrix/score_matrix-help.html). Using the Predictive Analytics Toolkit (PAT, 2017), a free Excel add-in that provides a point-and-click interface for conducting advanced prediction analysis from Excel using R packages, we then applied statistical and machine-learning algorithms for detecting, quantifying and visualizing dependencies between assay results and cancer classifications, including logistic regression and correlation, classification and regression trees (CARTs), Bayesian networks, and model ensembles (Random Forest ensembles of CART trees). This was followed by exploration of black-box predictive analytics algorithms, using the PAT, to maximize predictive power without considering interpretability of the underlying models. These statistical methods are described briefly below in Table 2; Full details of the methods are available in the Supplemental Material.

3. Results

The results of our statistical and prediction modeling analyses are briefly summarized in Table 2 for the dataset adjusted for the activity burst effect; results for the unadjusted dataset are presented in the Supplemental Material. Overall, the results indicate that, for the current ToxCast/Tox21 assays and datasets, predicting cancer based on data for the seven key characteristics, either alone or in combination, is fundamentally no better than chance—-whether or not adjustment is made for the activity burst phenomena.

3.1. Statistical analyses

Table 3 provides the summary descriptive statistics for assay hits per chemical over all the assays for the adjusted data, similar to the data representation in Hill et al. (2017). The distributions of assay hit calls for both positives and negatives are not normal and contain a high proportion of zeros. Similar to Hill et al. (2017), the Wilcoxon rank sum test was used to determine if the distributions for positives and negatives were statistically different—they were not for either cytotoxicity-adjusted data or unadjusted data. Using percent activity within each key characteristic complicates the choice of statistical tests because of the large numbers of zeros and ones and the ranking methodology used in the Wilcoxon test. Hence, the Kolmogorov-Smirnov (KS) two-sample test is likely the most appropriate test for determining any difference between the distributions of percent activity values for positives and negatives (Sokal and Rohlf, 1981).

The empirical cumulative distributions (CDF) for each of the seven key characteristics from the dataset adjusted for the activity burst phenomena are plotted in Fig. 2; these data were analyzed using the Kolmogorov-Smirnoff two sample test. None of the seven distribution pairs were significantly different for the dataset adjusted for the activity burst phenomena. The analysis of the ToxCast/Tox21 dataset unadjusted for the activity burst phenomena showed four of the distribution pairs were not significantly different, while the remaining three distributions pairs showed significantly greater ToxCast/Tox21 activity for chemicals classified as not having a human cancer hazard potential (see Supplemental Material). Similarly, a principal component analysis was also unable to distinguish positives from negatives over all seven key characteristics (Fig. 3; Supplemental Material for the data not adjusted for the burst phenomenon).

If the seven IARC key characteristics were predictive of carcinogenicity, then using another mapping technique, Jaccard similarity analysis, chemicals classified as having a human cancer hazard potential should segregate from the non-carcinogen/not likely carcinogens. However, as shown in Fig. 4, for the dataset adjusted for the activity burst phenomena, there is no readily apparent separation. Similar results were found for the dataset that was not adjusted for the burst activity (Supplemental Material).

3.2. Regression analysis and analysis using machine learning prediction algorithms

Table 4 presents the partial correlation coefficients obtained by correcting the correlation between each pair of variables for the levels of other variables using linear regression. For the dataset without adjustment for the burst phenomena, "Induces Oxidative Stress" was the only key characteristic significantly correlated with a positive carcinogenic hazard. However, in the analysis conducted on the dataset adjusted for the activity burst phenomena (e.g., cytotoxicity), this statistical significance ceased to exist (Table 4); none of the key characteristic were significantly correlated with having a positive cancer hazard potential at a p-value of <0.05. For the dataset adjusted for the activity burst phenomena, proliferation is a borderline significant predictor with a p value of 0.09.

Bayesian Network (BN) analysis (Fig. 5), where the arrows between any two nodes indicate a probability relationship of the two nodes, indicated a number of the key characteristics as significantly associated with one another for both the datasets. For example, for both datasets, the associations among "Oxidative Stress" and "Epigenetic Changes" and "Chronic Inflammation" were statistically significant. Nevertheless, similar to the linear regression analysis, for the dataset not adjusted for the activity burst phenomena, the BN analysis only identified "Induces Oxidative Stress" as the key characteristic significantly linked with a positive carcinogenic hazard (Fig. 5a), but this relationship ceased when the dataset that was adjusted for the activity burst phenomena was analyzed (Fig. 5b).

For both CART tree analysis and logistic regression, yet again,

Table 2
Summary of the analysis methods and results.

Descriptive Statistics of Assay Hit Calls per Chemical Across All Key Characteristics[a]
Purpose: Compare overall results across all assays in a fashion similar to that of Hill et al. (2017) to determine if substances classified as having human cancer hazard potential differed in ToxCast/Tox21 overall bioactivity from substances not posing a carcinogenic hazard.
Results: The Kolmogorov-Smirnov two sample test was used and no significant difference was found between these assay hit calls per chemical for positives and negatives in the cytotoxicity adjusted results with considerable certainty ($p = 0.9977$).

Kolmogorov-Smirnov Test of Empirical Cumulative Distributions Functions (CDFs)[a]
Purpose: Compare the CDFs (at the point of maximum discrepancy and the shapes of the CDFs) for each of the seven key characteristic to test the null hypothesis (i.e., to determine if the CDFs for the ToxCast data pairs for each key characteristic are statistically different).
Results: For the seven distribution pairs, the CDFs for the key characteristics for negatives were not different than the corresponding CDFs for positives. There was no meaningful difference between the ToxCast/Tox21 results for chemicals classified by EPA as having cancer hazard potential and substances EPA has determined do not pose a carcinogenic hazard. These results are highly certain (using $p < 0.05$ to determine a statistically significant difference, the calculated p values ranged from 0.33 to 1.0).

Principal Component Analysis (PCA)[a]
Purpose: PCA provides a means of visualizing multivariate data using a small number of factors normalizing to the mean value across the items being compared—in our case, two sets of chemicals. This method provides a means to compare variation of large datasets and visualize differences in the two datasets.
Results: The first three principal components with factor loadings based on the characteristics indicate that the PCA cannot clearly separate positives from negatives into distinct groups; no meaningful difference was seen between the ToxCast/Tox21 results for chemicals classified by EPA as having cancer hazard potential and substances EPA has determined do not pose a carcinogenic hazard. For the set of 248 substances considered here, the first three principal components accounted for almost 93% of the total variance in the data. On a 3D plot, however, no segregation of positives and negatives was observed.

Jaccard Similarity Analysis[a]
Purpose: To compare the similarity (and diversity) of the ToxCast results for each of the seven key characteristics of the negatives to the corresponding results for the positives. The Jaccard coefficient (ranging from 0 to 1), which measures similarity between pairs of chemicals based on the set of key characteristics. The Jaccard coefficient is similar to the Tanimoto metric used for cheminformatics similarity. For each pair of chemicals, the Jaccard coefficient is defined as the size of the intersection of the datasets divided by the size of the union of the datasets. If the 7 IARC key characteristics were predictive of carcinogenicity, then the positives should segregate from the negatives on the map (the Jaccard coefficient would approach 0).
Results: Similar to the results from PCA, there is no readily apparent separation of the positives and negatives; no meaningful difference was noted between the ToxCast/Tox21 results for chemicals classified by EPA as having cancer hazard potential and substances EPA has determined do not pose a carcinogenic hazard.

Logistic Regression and Correlation
Purpose: To measure the relationship between the categorical dependent variable (ToxCast results for chemicals classified by EPA as having cancer hazard potential) and the key characteristics as independent variables by estimating probabilities using a logistic function. Partial correlation coefficients were obtained by correcting the correlation between each pair of variables for the levels of other variables using linear regression.
Results: None of the key characteristic were significantly correlated with the chemicals classified by EPA as having cancer hazard potential. These results are highly certain (using $p < 0.05$ to determine a statistically significant difference, the calculated p values ranged from 0.09 to 1.0).

Classification and Regression Trees (CARTs)
Purpose: Machine-learning methods were used for constructing prediction models from ToxCast data. The models were developed using decision trees based upon the ToxCast data for the seven key characteristics of the positives and negatives. The tree structure is applicable to any number of variables, and models are evaluated in terms of their ability to predict the class to which the data belongs.
Results: The CART tree analysis for the dependent variable (i.e. chemicals classified by EPA as having cancer hazard potential) did not identify any of the seven key characteristics as a significant predictor of being classified as a carcinogen.

Bayesian Network Analysis
Purpose: To evaluate the probabilistic relationships between the ToxCast data and the seven key characteristics of carcinogens. All seven key characteristics and ToxCast data were assembled into a conditional probability table and machine learning algorithms were used to perform inference and learning in Bayesian Networks (BN) to characterize the maximum a posterori probability that ToxCast results can predict carcinogenicity.
Results: Although a number of the key characteristics were seen to be significantly correlated with one another for both the datasets. For the BN models developed from the ToxCast/Tox21 data for the seven key characteristics were not able to predict carcinogenicity. These results are highly certain at a p-value of <0.05.

Model Ensembles (Random Forest Ensembles of CART Trees)
Purpose: Employ Random Forests analysis (an algorithm and computational procedures) to investigate and discover complex nonlinear relationships between explanatory (ToxCast data for each of the seven key characteristics) and outcome variables (chemicals classified by EPA as having cancer hazard potential and substances EPA has determined do not pose a carcinogenic hazard).
Results: No predictive relationships were observed, with the exception that there was a statistically significant association, threshold-like nonlinearity, between proliferation and the probability of being a chemical classified by EPA as having cancer hazard potential (based on Spearman's rank correlation of 0.54 and p-value 0.00)

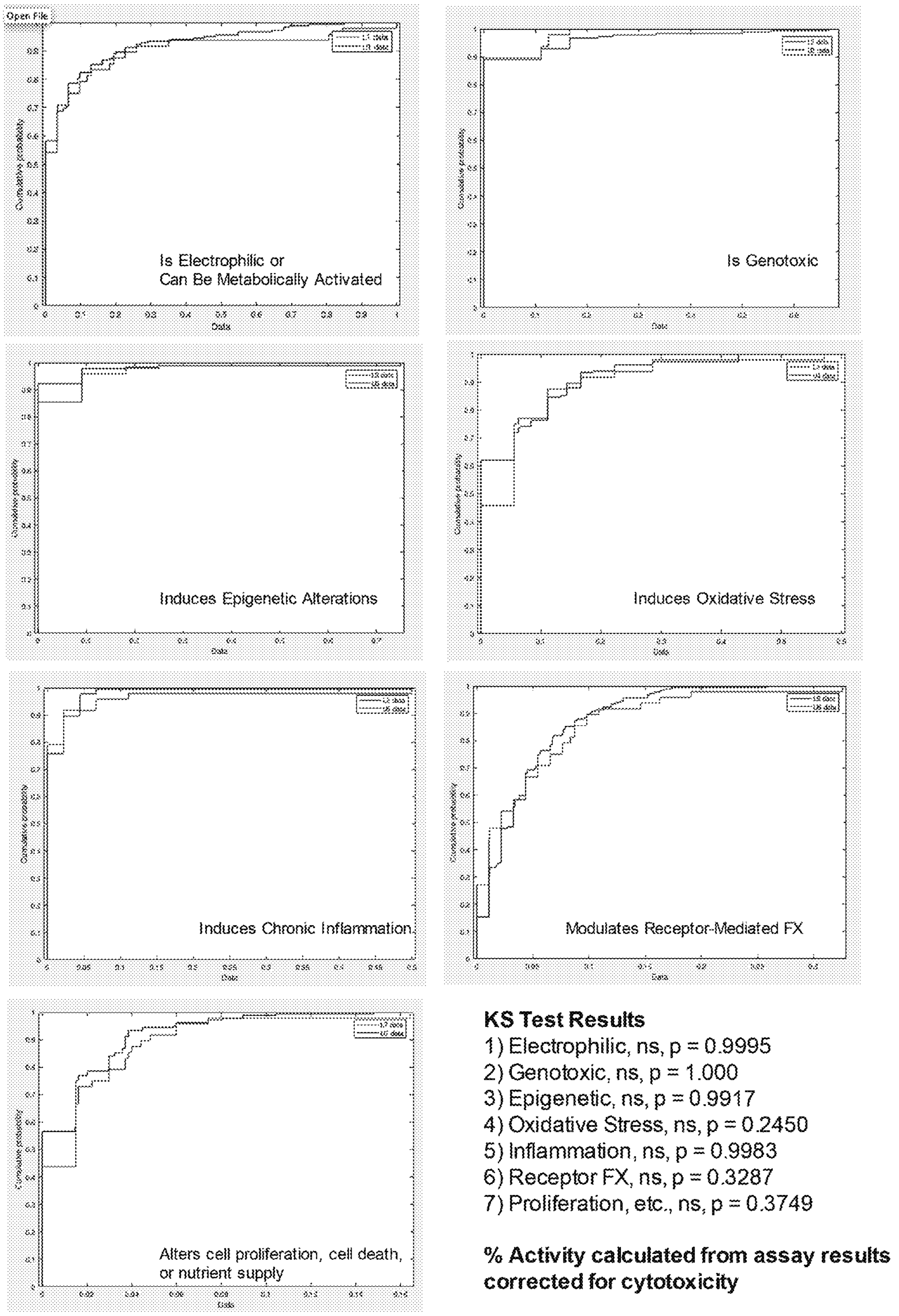
Black-Box (BB) Classification and Regression Predictive Modeling
Purpose: Using machine learning to develop predictive mathematical models based on the EPA classifications of carcinogenic potential and the ToxCast data and the seven key characteristics of carcinogens. Disjoint training and test subsets of the data were automatically created and multiple models (including CART trees and random forest ensembles) were fit to the training data and then evaluated on the test set.
Results: None of these automated predictive modeling algorithms performed well on the test set. The BB models developed from the ToxCast/Tox21 data for the seven key characteristics were not able to predict carcinogenicity.

[a] Because not all assays in each characteristic were tested for each chemical, the dataset necessarily contained nulls. For the KS test, PCA and Jaccard similarities, the methods were adjusted specifically to take into account the presence of null data.

Table 3
Summary descriptive statistics for assay hit calls for positives and negatives from the dataset adjusted for the activity burst phenomenon.

Assay Hits Per Chemical	Mean	Standard Deviation	Count	Median
Positives	27.4	22.7	54	24.5
Negatives	19.7	20	194	11.5

without adjustment for the activity burst phenomena, oxidative stress was the only significant predictor (Table 5) and only at relatively high activity levels (>0.438), which are attained by only 9 of 248 chemicals (Supplemental Material) and for the dataset that was adjusted to account for the activity burst phenomena, none of the seven key characteristics were identified as a predictor of carcinogenicity (Table 5). A similar pattern was observed with Random Forest analysis (an ensemble of CART trees fit to random samples of the full data set). For the dataset adjusted for activity burst phenomena, a partial dependence plot revealed a threshold-like nonlinearity between oxidative stress and probability of posing a carcinogenic hazard. After adjustment for the activity burst effect, this relationship with oxidative stress ceased, and a statistically significant threshold-like nonlinearity between proliferation and the probability of posing a carcinogenic hazard was observed (see Supplemental Material for details). Consistent with the foregoing



KS Test Results

1) Electrophilic, ns, p = 0.9995
2) Genotoxic, ns, p = 1.000
3) Epigenetic, ns, p = 0.9917
4) Oxidative Stress, ns, p = 0.2450
5) Inflammation, ns, p = 0.9983
6) Receptor FX, ns, p = 0.3287
7) Proliferation, etc., ns, p = 0.3749

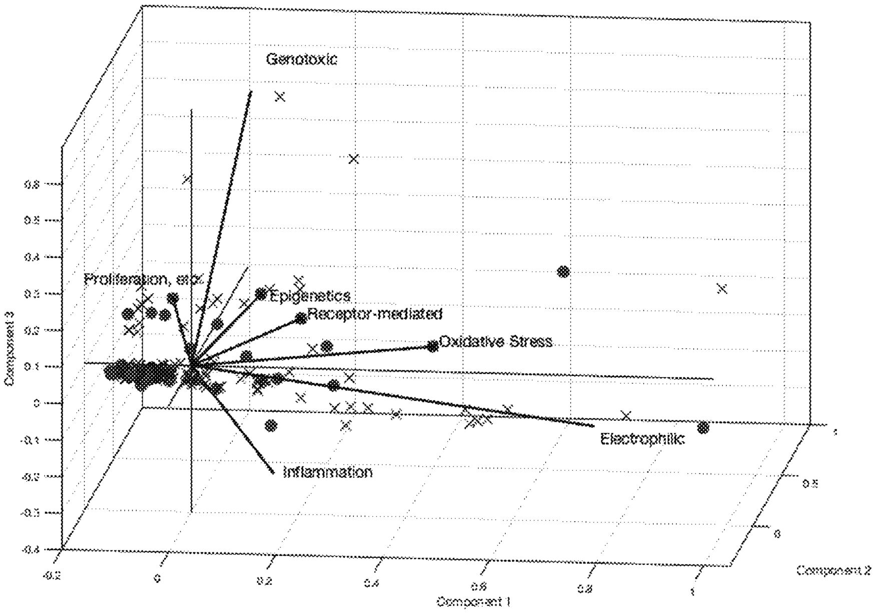
% Activity calculated from assay results corrected for cytotoxicity

Fig. 2. Empirical cumulative distributions of percent activity for chemicals classified as having a human cancer hazard potential (designated L and plotted in red) and chemicals classified as having a human cancer hazard potential from those that do not pose a carcinogenic hazard (designated UL and plotted in blue) for each of seven key characteristics from the dataset that was adjusted for the activity burst effect. The list in the lower right shows the p-values for the Kolmorgorov-Smirnov two-sample test (ns: no significant difference).

results, using black-box predictive analytics, in which disjoint training and test subsets of the data were automatically created and multiple models (including CART trees and random forest ensembles) were fit to the training data and then evaluated on the test set, none of these automated predictive modeling algorithms performed well on the test set, with or without adjustment for the activity burst effect (see Supplemental Material for details).

4. Discussion

A great deal of research attention has focused on establishing confidence in HTS assays (such as ToxCast/Tox21) as markers for complex apical biological responses such as chemical carcinogenesis. Some scientists passionately champion, some merely hope for, whilst others decry the idea that these HTS assay outputs may



3D plot of PCA conducted with cytotoxicity-adjusted % activity. There is no clear separation of likely carcinogens (red circles) and chemicals not likely to be carcinogens (blue x's). The explained variances for the seven principle components are:
1 Electrophilic (44.0%)
2 Genotoxic (33.9%)
3 Epigenetic (8.5%)
4 Oxidative Stress (7.8%)
5 Inflammation (3.5%)
6 Receptor FX (1.3%)
7 Proliferation, etc. (1.0%)

Fig. 3. Three dimensional plot of the Principal Component Analysis (PCA) results including all seven key characteristics for the chemicals classified as posing carcinogenic hazard (red filled circles) and chemicals that do not pose a carcinogenic hazard (blue x-marks) from the dataset that was adjusted for the activity burst effect.

facilitate more rapid and economical identification of chemical toxicities. Scientific datasets from ToxCast/Tox21 HTS assays along with existing evaluations of chemical cancer hazards offer unique opportunities to develop and evaluate hypotheses related to the use of HTS data for hazard characterization.

4.1. ToxCast/Tox21 data cannot reliably distinguish or predict chemicals that pose a carcinogenic hazard

Our extensive analysis stands in contrast to the implied inference model, or supposition, that ToxCast/Tox21 assay results are an indication of biological activity underpinning mechanisms that can be causally related to classifying potential carcinogenic hazards to humans. We found no statistically significant differences in ToxCast/Tox21 responses for the seven IARC key characteristics of carcinogens between "positives"—substances determined to have human cancer hazard potential—and "negatives"—substances that pose no cancer hazard. The only exception was for the ToxCast/Tox21 dataset unadjusted for cytotoxicity; for that dataset, oxidative stress was the sole predictor in a few of the analyses and the predictive power was marginal. However, after adjusting for cytotoxicity, the predictiveness of this key characteristic ceased. Overall, for this dataset, we cannot reject the null hypothesis that the seven IARC key characteristics represented in ToxCast/Tox21 assay results fail to differentiate substances USEPA classifies as having cancer hazard potential from substances not posing a carcinogenic hazard.

To date, taken collectively, most results (and models) indicate that ToxCast/Tox21 assays fall short in reliably predicting human cancer risk classifications. The results reported herein are consistent with the previous reports of Thomas et al. (2012), Benigni (2013, 2014), Cox et al. (2016) and the recent report from Hill et al. (2017). Similar to our study, Hill et al. (2017) also used ToxCast assay hit-calls and cancer classifications from USEPA's OPP to evaluate the ability of HTS results to predict cancer. Hill et al., 2017 did not aggregate results for characteristics of carcinogens, but rather grouped the ToxCast HTS assays into three categories: all assays, all cancer-related assays as identified by Kleinstreuer et al. (2013), and assays that targeted cytotoxicity as an endpoint. Using ToxCast data that excluded activity within the burst region, Hill et al. (2017) demonstrated that neither all assays (HTS hits per chemical) nor HTS results for cancer-related assays were capable of distinguishing USEPA's "probable/likely" carcinogens from "not likely" carcinogens. Only one comparison reported by Hill and colleagues was significant—for the HTS cytotoxicity category, the number of hits per chemical was greater for the substances deemed by USEPA as "probable or likely" carcinogens. Moreover, Hill et al. (2017) confirmed the determination of Cox et al. (2016) that the prediction modeling reported by Kleinstreuer et al. (2013) correlating cancer pathway bioactivity scores based on ToxCast data with *in vivo* carcinogenic effects in rodents was questionable, and, for the most part, is no better than chance alone.

In the dataset lacking the adjustment for the burst phenomenon, the predictive power of oxidative stress was statistically significant in a limited number of analyses; these relationships,

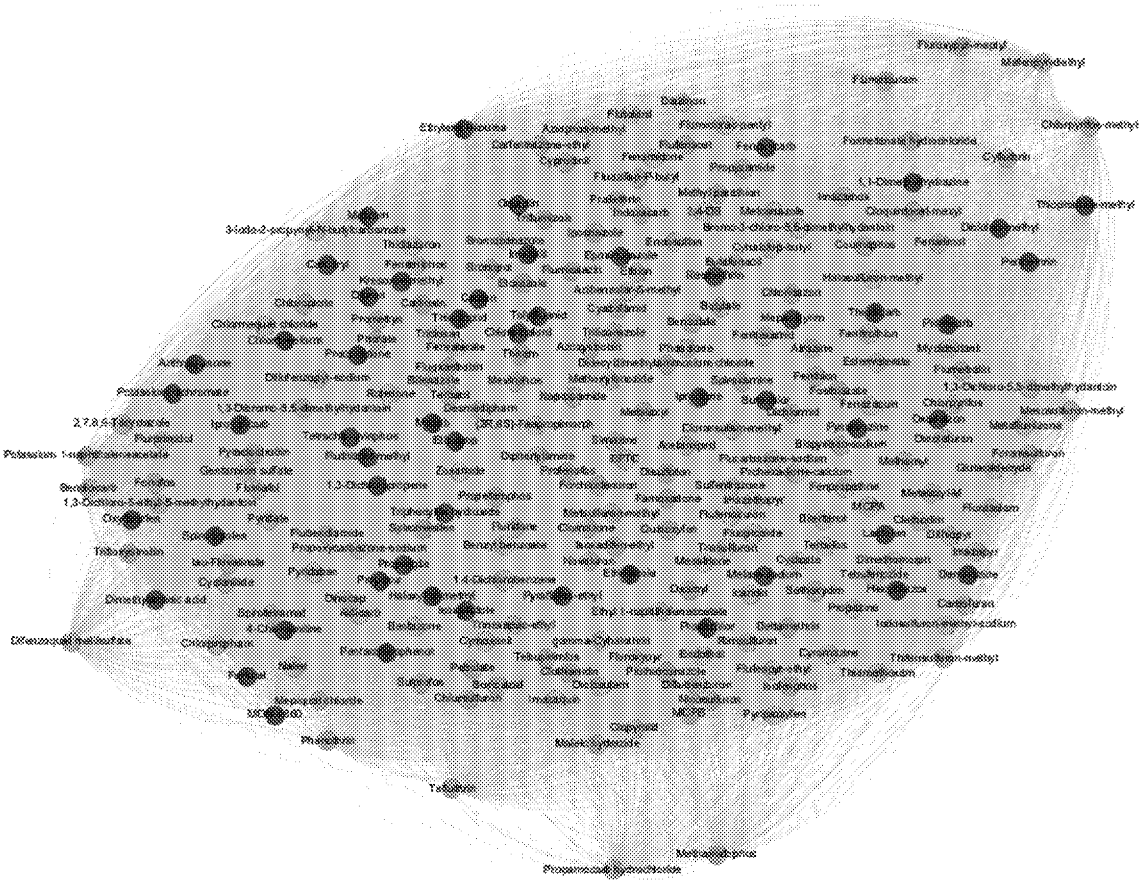


Fig. 4. Network visualization map of chemicals classified as posing carcinogenic hazard (red) and those that do not pose a carcinogenic hazard (blue) based on the similarity of hit call pattern of each chemical for the seven key characteristics from the dataset that was adjusted for the activity burst effect. The 54 positives are interspersed with the 194 negatives; whereas if the key characteristics provided a means of distinguishing carcinogens from non-carcinogens, segregation of positives and negatives over the two-dimensional map space would be expected.

Table 4
Partial Correlations between the key characteristic and whether a chemical is a likely carcinogen (i.e., chemical classified as posing carcinogenic hazard) or unlikely carcinogen (i.e., chemical that does not pose a carcinogenic hazard).

Partial Correlations of Key Characteristics With "Likely" (classified as having cancer hazard potential)		
Characteristic	Partial Correlation value (p value)	
	Dataset adjusted for activity burst phenomena	Dataset not adjusted for activity burst phenomena
Likely	1.000 (0.00)	1.000 (0.00)
Electrophilic	0.0635 (0.34)	−0.116 (0.07)
Genotoxic	−0.0929 (0.17)	−0.031 (0.63)
Epigenetic	−0.0249 (0.71)	0.041 (0.53)
Oxidative	−0.0041(0.95)	0.136 (0.03)*
Inflammation	0.0569 (0.40)	0.024 (0.71)
Receptor	0.0003 (1.0)	−0.014 (0.83)
Proliferation	0.1128 (0.09)	0.077 (0.23)

*significant <0.05.

however, ceased when the dataset was adjusted for the burst phenomenon. These findings indicate considerable caution needs to be exercised when evaluating ToxCast/Tox21 results for assays determined to be relevant to the key characteristic of oxidative stress. Bus (2016) analyzes and discusses in detail the challenges in the use of oxidative stress as a key characteristic and documents limitations of the IARC cancer hazard evaluation of glyphosate that used published studies to infer strong evidence of oxidative stress. For example, the IARC Working Group examining mechanistic data for glyphosate ignored toxicokinetics and uncritically relied on *in vitro* studies in which the test concentrations could not have been physically attained in any *in vivo* animal test, much less humans. Although the reliance on ToxCast/Tox21 oxidative stress results by the IARC Working Groups engaged in Monographs 112 and 113 appears to have been limited, the fact that the data was not adjusted for the burst activity effect should be noted. In the future, given the relationship between high-dose cytotoxicity, the activity burst phenomenon and ToxCast/Tox21 assays deemed indicative of the characteristic of oxidative stress, it is imperative that adjustment for the burst phenomenon be conducted in advance of interpreting results as supporting evidence of activity relevant to a cancer hazard for humans.

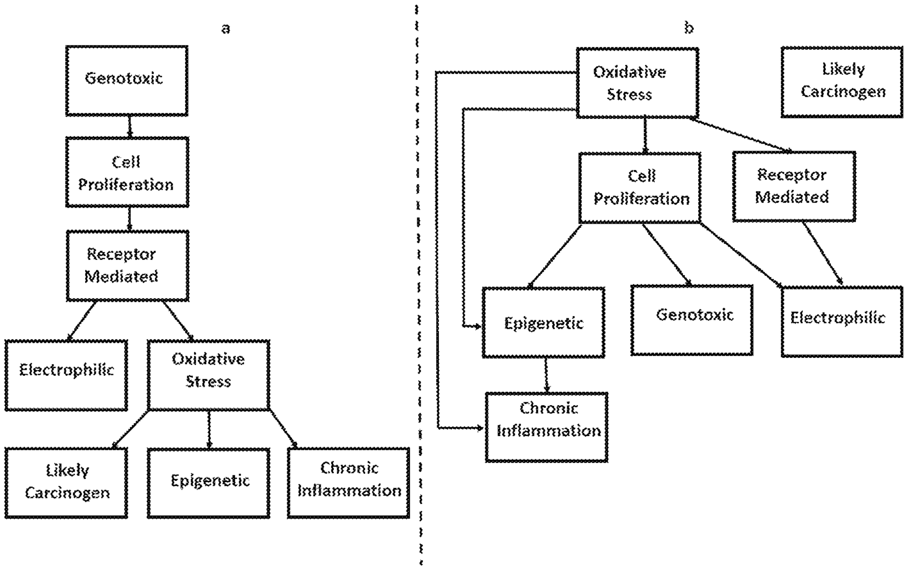


Fig. 5. Bayesian Network developed from percent activity values for the seven key characteristics. Fig. 4a is the BN for the dataset that was not adjusted for the activity burst effect and 4b is the BN for the dataset that was adjusted for the activity burst effect.

Table 5
Logistic regression results.

	Central Estimate	Standard Error	z Value	Pr (>|z|)	Central Estimate	Standard Error	z Value	Pr (>|z|)
	Dataset not adjusted for activity burst phenomena				Dataset adjusted for activity burst phenomena			
(Intercept)	−1.6373	0.2722	−6.0	2.00E-09**	−1.56	0.24	−6.5	8e-11 **
Electrophilic	−1.0140	0.5566	−1.8	0.07	0.90	0.99	0.9	0.4
Genotoxic	−0.4938	0.9866	−0.5	0.62	−3.77	3.10	−1.2	0.2
Epigenetic	0.4011	0.6919	0.6	0.56	−0.84	2.39	−0.4	0.7
Oxidative Stress	3.1672	1.5714	2.0	0.04*	−0.17	2.71	−0.1	1.0
Inflammation	1.1101	2.8550	0.4	0.70	2.30	3.10	0.7	0.5
Receptor Mediated	0.0011	2.6923	0.0	1.00	0.36	5.12	0.1	0.9
Proliferation	2.3721	1.9643	1.2	0.23	12.73	7.87	1.6	0.1

**significant < 0.001 *significant <0.05.

Similarly, the ToxCast assays mapped to the characteristic "Is Genotoxic" comprise only 9 assays, all of which are P53 constructs. Although these p53 ToxCast/Tox21 data were used in IARC Monograph 112, in the subsequent Monograph 113, the p53 assay results and the characteristic "Is Genotoxic" was dropped from the ToxCast/Tox21 section of the analysis; only ToxCast data mapped to the remaining six key characteristics were used in this section. Although Monograph 113 provided no explanation for the statement "no assay [ToxCast/Tox21] end-points were mapped to this [Is Genotoxic] characteristic," it seems likely that the IARC Working Group for Monograph 113 concluded that 9 assays all focusing on P53 did not robustly represent this key characteristic.

Goodman and Lynch (2017) have criticized the IARC processes for integrating mechanistic evidence used by the Working Groups as lacking rigor and transparency. In conducting this analysis, we confirmed that the assignments of various Tox21/ToxCast assays to key characteristics, the process IARC follows for integrating ToxCast/Tox21 data (and/or other mechanistic evidence), and the criteria for characterizing the totality of mechanistic evidence as "strong," "moderate," "weak," or "inadequate" for each characteristic have not been explicitly documented and apparently have not been subjected to independent scientific peer review/publication in the open scientific literature. This lack of transparency and scientific rigor, taken together with our findings that ToxCast/Tox21 results for the seven IARC key characteristics fail to differentiate carcinogens from non-carcinogens, raises doubts about the proposed use of such mechanistic evidence for elevating human cancer hazard classifications (Guyton, 2015).

4.2. Challenges Using ToxCast/Tox21 data and key characteristics of carcinogens

An important limitation of HTS-based prediction modeling stems from the fact that the ToxCast/Tox21 assays themselves have not, for the most part, been specifically designed to evaluate key steps or stages of the pathogenesis of chemically-induced cancer. Smith et al. (2016) reported that, at the initial 2012 IARC workshop, invited experts identified twenty four mechanistic endpoints, each with a number of subcategories, and then at a subsequent workshop, the invited participants merged a number of the characteristics to arrive at the ten key characteristics. Although not evaluated by IARC or here, for each key characteristic, the extent of variability of results within and across ToxCast/Tox21 assays, as well as the sensitivity, specificity and domain of applicability of each ToxCast/Tox21 assay, are important elements to take into account for both implied and explicit inference. Even though the IARC Working Groups were presumably able to assign ToxCast/Tox21 assays to the key characteristics, a number of questions have yet to be fully addressed, such as: How well do the assigned assays reflect the underlying biology or mode of action? How well do the key characteristics actually represent cancer pathogenesis? What set of key characteristics and assays are optimal?

Although Hanahan and Weinberg (2000, 2011) provide considerable insight into the nature of cancer pathogenesis by identifying biological hallmarks, the authors emphasize that identification and confirmation of the hallmarks is still a work in progress. For example, inflammatory response is identified as an emerging or enabling characteristic by Hanahan and Weinberg (2011) and Smith et al. (2016) identify the ability to induce inflammation as a characteristic of carcinogens, but whether inflammation is a causative factor for later hallmarks or a result of earlier hallmarks remains an open question.

A limitation of this analysis is that only ToxCast/Tox21 data were considered, whereas in the recent IARC evaluations, results from other mechanistic assays have apparently been included. Nevertheless, the work described here and other studies demonstrate that the current battery of HTS assays does not yet produce data that can reliably predict carcinogenic hazard (Thomas et al., 2012; Cox et al., 2016; Hill et al., 2017). As the understanding of cancer MOAs progresses and new assay technologies are developed, endpoint-specific suites of assays and computational methods that reflect both toxicokinetics and cellular pathways may allow for better predictive performance. At the present time, one cannot conclude from *in vitro* studies alone using assays currently mapped to characteristics of carcinogens that one or multiple molecular mechanisms are likely to be operative in inducing cancer or creating a cancer hazard. At most, the ToxCast/Tox21 data can indicate the potential for a chemical to interact with one or more biological pathways. But interaction at a molecular or cellular level is not the same as causation of an adverse effect such as cancer *in vivo*.

A common goal of the ToxCast/Tox21 program and the Adverse Outcome Pathway (AOP) effort is to discover which specific assays reflect initial molecular events or early key events within biological pathways (LaLone et al., 2017; van Bilsen et al., 2017). While many of the HTS assays have been designed to be exquisitely sensitive, these assays are uncoupled from the normal physiological networks that occur *in vivo* and therefore lack appropriate biological context (e.g. Simon et al., 2015). Such cellular pathways and networks involved in homeostasis can have “clear-cut, mechanistically definable thresholds” or tipping points (Zhang et al., 2014). The concept of “molecular tipping points” and the idea that detection of bioactivity within early key events may not necessarily lead to the manifestation of an adverse outcome is made abundantly clear in a comprehensive analysis of the MOA of phthalate-induced liver tumors in mice by activation of PPARα (Lake et al., 2016). The concept of dose-dependent tipping points is intrinsic to key event relationships and necessary to fully understand the role of these early key events within pathways of cancer pathogenesis (Shah et al., 2016). Use of binary hit calls to encode assay results instead of expressions of potency and efficacy ignores both the idea of “tipping points” and what other later obligatory key events must occur for cancer progression. While the hallmarks of cancer are reflective of events within both normal homeostatic mechanisms and cancer pathogenesis, the use of binary hit calls as an indicator of relevant activity in such pathways is likely too generalized as it ignores important differences in dose and potency, and dose-dependent transitions (Slikker et al., 2004) that distinguish normal physiology from pathological processes. Thus, approaches such as those of IARC which appear to use mechanistic evidence, including ToxCast/Tox21 data, are flawed and misleading because they do not explicitly integrate dosimetry, temporality, and causality inherent in key event relationships.

4.3. Opportunities to address the challenges

Our results should not be interpreted to suggest that mechanistic data have no role to play in informing determinations of potential carcinogenic risks of chemicals to human. Rather, the findings demonstrate an urgent need for explicit, transparent, and scientifically robust procedures to evaluate the relevance and reliability of mechanistic datasets and the process of integrating mechanistic results with extant animal toxicity findings and human epidemiology. Goodman and Lynch (2017) reached a similar conclusion, recommending that IARC develop and apply “explicit guidance for how to consider the totality of the mechanistic evidence, including study strengths and limitations .... ,” noting that “[A]dopting a systematic approach for evaluating and integrating mechanistic evidence with the other realms of evidence will allow for hazard classifications that are scientifically defensible and appropriate for regulatory decision-making.”

The scientific confidence framework (Cox et al., 2014; Patlewicz et al., 2015 and Cox et al., 2016) designed to aid in developing, evaluating and communicating the scientific confidence in Tox21 assays and their prediction models was derived in part from guidance on the use of biomarkers in medicine (IOM, 2010). Adoption of such a framework should be considered to enhance the rigor and transparency of the IARC process for integration of mechanistic evidence. The elements of this scientific confidence framework are: (1) Analytical validation of mechanistic assays that includes documentation of the biological basis and analytical performance of assays, including reliability, sensitivity, specificity and domain of applicability for all assays; (2) Qualification of inference or prediction models based on mechanistic assays, a transparent characterization that includes explicit decision criteria, decision logic or a defined algorithm for each evidence integration model, appropriate measures of goodness-of-fit, robustness and performance documentation in sufficient detail to facilitate review, reconstruction and independent verification and replication of results; and (3) Utilization, a specification of the confidence that supports the fitness inference decision criteria, decision logic or prediction model derived from mechanistic assays for a specific purpose or use. For all three aspects of this framework, the scientific justification underpinning the use for a specific decision context requires documentation with sufficient detail to enable an independent scientific reviewer to replicate the analysis.

Another reason the IARC approach falls short is that it does not explicitly incorporate understanding of the causal linkages of the sequence of key events and biological responses (including dose-response and temporal relationships) involved in carcinogenesis. For incorporating mechanistic data into cancer hazard evaluations, we specifically recommend adoption of the AOP (OECD, 2016) or MOA framework (Meek et al., 2014) that articulates toxicity pathways comprised of sequences of key events, starting with an initial molecular event, followed by a series of key events linked to one another, ultimately resulting in a specific adverse outcome (Meek et al., 2013, 2014). Mechanistic and observational datasets can then be organized and aligned with corresponding key events and key event relationships, and defined Bradford Hill causal considerations can be applied to assess biological plausibility, essentiality and empirical evidence (Becker et al., 2017; Meek et al., 2013, 2014; OECD, 2016). Furthermore, adoption of the recently developed causal analysis scoring approach for evaluating the weight of evidence of potential cancer MOAs (Becker et al., 2017), built from the WHO/IPCS MOA framework, would provide a systematic and transparent approach for evaluating a chemical dataset, including mechanistic data, using hypothesized MOAs and the evolved Bradford Hill causal considerations for integrating evidence. This quantitative confidence scoring approach facilitates side by side comparison of different hypothesized cancer MOAs and is useful in characterizing the likely operative MOA.

5. Note added in proof

While this manuscript was undergoing peer review, Chiu et al. (2017) published "Use of high-throughput *in vitro* toxicity screening data in cancer hazard evaluations by IARC Monograph Working Groups" (ALTEX (online first, http://dx.doi.org/10.14573/altex.1703231). Unlike the investigation we conducted, the Chiu et al. (2017) study did not include adjustment of ToxCast/Tox21 results for the activity burst phenomena. As shown in our analysis, lack of consideration of this burst effect (e.g., failure to account for cytotoxicity) leads to scientifically questionable, and at times, unfounded, conclusions. Furthermore, Chiu et al. (2017) did not evaluate the ability of HTS data to predict cancer classifications. Instead, Chiu and colleagues employed a case study approach that appears to essentially mirror the methodology used in the recent IARC Monographs; they relied on ToxPi evaluations for activity rankings, HTS activity hit frequencies and an expert judgment approach (without accompanying *a priori* science-based ground rules or systematic guidance) to integrate mechanistic data (including not only ToxCast/Tox21 data but apparently also results from other *in vitro* and *in vivo* mechanistic studies). We recommend that readers carefully evaluate the strengths and limitations of our study, and those of Chiu et al. (2017) in determining the evidentiary value of HTS ToxCast/Tox21 results for informing decisions on potential carcinogenic risks of chemicals.

Acknowledgments

The authors thank N. Ryan, D. Juberg and D. Geter for their thoughtful review of an early draft of this manuscript. In addition, we sincerely appreciate the thoughtful and very constructive comments provided by the independent peer reviewers. The authors had complete control over the design, conduct, interpretation, and reporting of the analyses included in this manuscript. The contents of this manuscript are solely the responsibility of the authors and do not necessarily reflect the views or policies of their employers. Support for this research was provided by the American Chemistry Council. R. Becker is employed by ACC, a trade association of U.S. chemical manufacturers.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.yrtph.2017.08.021.

Transparency document

Transparency document related to this article can be found online at http://dx.doi.org/10.1016/j.yrtph.2017.08.021.

References

Becker, R.A., Dellarco, V., Seed, J., Kronenberg, J., Meek, M.E., Foreman, J., et al., 2017. Quantitative weight of evidence to assess confidence in alternative modes of action. Regul. Toxicol. Pharmacol. 86, 205–220.

Benigni, R., 2013. Evaluation of the toxicity forecasting capability of EPA's ToxCast phase I data: can ToxCast *in vitro* assays predict carcinogenicity? J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev. 31, 201–212.

Benigni, R., 2014. Predicting the carcinogenicity of chemicals with alternative approaches: recent advances. Expert Opin. Drug Metab. Toxicol. 10, 1199–1208.

Boobis, A.R., Cohen, S.M., Dellarco, V., McGregor, D., Meek, M.E., Vickers, C., Willcocks, D., Farland, W., 2006. IPCS framework for analyzing the relevance of a cancer mode of action for humans. Crit. Rev. Toxicol. 36, 781–792.

Boobis, A.R., Cohen, S.M., Dellarco, V.L., Doe, J.E., Fenner-Crisp, P.A., Moretto, A., Pastoor, T.P., Schoeny, R.S., Seed, J.G., Wolf, D.C., 2016. Classification schemes for carcinogenicity based on hazard-identification have become outmoded and serve neither science nor society. Regul. Toxicol. Pharmacol. 82, 158–166.

Bus, J.S., 2016. IARC use of "oxidant stress" as key mode of action characteristic for facilitating cancer classification: glyphosate case example illustrating of a lack of robustness in interpretative implementation. Regul. Toxicol. Pharmacol. 86, 157–166.

Chiu, W.A., Guyton, K.Z., Martin, M.T., Reif, D.M., Rusyn, I., 2017 Jul 24. Use of high-throughput in vitro toxicity screening data in cancer hazard evaluations by IARC Monograph Working Groups. ALTEX. http://dx.doi.org/10.14573/altex.1703231.

Cox, L.A., Popken, D., Marty, M.S., Rowlands, J.C., Patlewicz, G., Goyak, K.O., Becker, R.A., 2014. Developing scientific confidence in HTS-derived prediction models: lessons learned from an endocrine case study. Regul. Toxicol. Pharmacol. 69, 443–450.

Cox, L.A., Popken, D.A., Kaplan, A.M., Plunkett, L.M., Becker, R.A., 2016. How well can in vitro data predict *in vivo* effects of chemicals? Rodent carcinogenicity as a case study. Regul. Toxicol. Pharmacol. 77, 54–64.

Dourson, M., Becker, R.A., Haber, L.T., Pottenger, L.H., Bredfeldt, T., Fenner-Crisp, P., 2013. Advancing human health risk assessment: integrating Recent Advisory Committee recommendations. Crit. Rev. Toxicol. 43, 467–492.

Goodman, J., Lynch, H., 2017. Improving the international agency for research on Cancer's consideration of mechanistic evidence. Toxicol. Appl. Pharmacol. 319, 39–46.

Guyton, K.Z., 2015. Systematic Identification of the Mechanistic Evidence for Cancer Hazard Assessment: Experience of the IARC Monographs Programme. http://ofmpub.epa.gov/eims/eimscomm.getfile?p_download_id=526753. (Accessed 24 August 2017).

Hanahan, D., Weinberg, R.A., 2000. The hallmarks of cancer. Cell 100, 57–70.

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. Cell 144, 646–674.

Hill 3rd, T., Nelms, M.D., Edwards, S.W., Martin, M., Judson, R., Corton, J.C., Wood, C.E., 2017. Editor's highlight: negative predictors of carcinogenicity for environmental chemicals. Toxicol. Sci. 155, 157–169.

IARC (International Agency for Research on Cancer), 2017. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, me 112. Some Organophosphate Insecticides and Herbicides: Diazinon, Glyphosate, Malathion, Parathion, and Tetrachlorvinphos. http://monographs.iarc.fr/ENG/Monographs/vol112/index.php. (Accessed 24 August 2017). and Annex 1 Supplementary Material to Volume 112. http://monographs.iarc.fr/ENG/Monographs/vol112/112-Section4-Spreadsheet.xlsx [accessed 24 August 2017].

IARC, 2016a. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, vol. 110. Some Chemicals Used as Solvents and in Polymer Manufacture. http://monographs.iarc.fr/ENG/Monographs/vol110/index.php. (Accessed 24 August 2017).

IARC, 2016b. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, vol. 113, 2,4-D and Some Organochlorine Insecticides. http://monographs.iarc.fr/ENG/Monographs/vol113/index.php. (Accessed 24 August 2017).

IARC, 2016c. Preamble. http://monographs.iarc.fr/ENG/Preamble/CurrentPreamble.pdf. (Accessed 24 August 2017).

Institute of Medicine (IOM), 2010. Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease.

Judson, R., Houck, K., Martin, M., Richard, A.M., Knudsen, T.B., Shah, I., et al., 2016. Editor's highlight: analysis of the effects of cell stress and cytotoxicity on in vitro assay activity across a diverse chemical and assay space. Toxicol. Sci. vol. 152, 323–339 (and Erratum at Toxicol Sci 153: 409).

Kleinstreuer, N.C., Dix, D.J., Houck, K.A., Kavlock, R.J., Knudsen, T.B., Martin, M.T., et al., 2013. *In vitro* perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis. Toxicol. Sci. 131, 40–55.

Lake, A.D., Wood, C.E., Bhat, V.S., Chorley, B.N., Carswell, G.K., Sey, Y.M., Kenyon, E.M., Padnos, B., Moore, T.M., Tennant, A.H., Schmid, J.E., George, B.J., Ross, D.G., Hughes, M.F., Corton, J.C., Simmons, J.E., McQueen, C.A., Hester, S.D., 2016. Dose and effect thresholds for early key events in a PPARalpha-mediated mode of action. Toxicol. Sci. 149, 312–325.

LaLone, C.A., Ankley, G.T., Belanger, S.E., Embry, M.R., Hodges, G., Knapen, D., Munn, S., Perkins, E.J., Rudd, M.A., Villeneuve, D.L., Whelan, M., Willett, C., Zhang, X., Hecker, M., 2017. Advancing the adverse outcome pathway framework-An international horizon scanning approach. Environ. Toxicol. Chem. 36, 1411–1421.

Lauby-Secretan, B., Loomis, D., Grosse, Y., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., et al., 2013. Carcinogenicity of polychlorinated biphenyls and polybrominated biphenyls. Lancet Oncol1 4, 287–288.

Lauby-Secretan, B., Loomis, D., Baan, R., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., et al., 2016. Use of mechanistic data in the IARC evaluations of the carcinogenicity of polychlorinated biphenyls and related compounds. Environ. Sci. Pollut. Res. Int. 23, 2220–2229.

Loomis, D., Guyton, K., Grosse, Y., El Ghissasi, F., Bouvard, V., Benbrahim-Tallaa, L., et al., 2015. Carcinogenicity of lindane, DDT, and 2,4-dichlorophenoxyacetic acid. Lancet Oncol. 16, 891–892.

McHale, C.M., Zhang, L., Smith, M.T., 2012. Current understanding of the mechanism of benzene-induced leukemia in humans: implications for risk assessment. Carcinogenesis 33, 240–252.

Meek, M.E., Boobis, A., Cote, I., Dellarco, V., Fotakis, G., Munn, S., et al., 2013. New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis. J. Appl. Toxicol. 34, 1–18.

Meek, M.E., Palermo, C.M., Bachman, A.N., North, C.M., Jeffrey Lewis, R., 2014. Mode of action human relevance (species concordance) framework: evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. J. Appl. Toxicol. 34, 595–606.

OECD, 2016. Users' Handbook Supplement to the Guidance Document for Developing and Assessing Adverse Outcome Pathways. OECD Series on Adverse

Outcome Pathways, No. 1. OECD publishing, Paris. Available at: http://www.oecd-ilibrary.org/environment/users-handbook-supplement-to-the-guidance-document-for-developing-and-assessing-adverse-outcome-pathways_5jlv1m9d1g32-en. (Accessed 24 August 2017).

Patlewicz, G., Simon, T.W., Rowlands, J.C., Budinsky, R.A., Becker, R.A., 2015. Proposing a scientific confidence framework to help support the application of adverse outcome pathways for regulatory purposes. Regul. Toxicol. Pharmacol. 71, 463–477.

Predictive Analytics Toolkit (PAT). 2017. http://cox-associates.com/patkit/[accessed 15 May 2017].

Shah, I., Setzer, R.W., Jack, J., Houck, K.A., Judson, R.S., Knudsen, T.B., et al., 2016. Using ToxCast data to reconstruct dynamic cell state trajectories and estimate toxicological points of departure. Environ. Health Perspect. 124, 910–919.

Simon, T.W., Budinsky, R.A., Rowlands, J.C., 2015. A model for aryl hydrocarbon receptor-activated gene expression shows potency and efficacy changes and predicts squelching due to competition for transcription co-activators. PLoS One 10 (6), e0127952. http://dx.doi.org/10.1371/journal.pone.0127952 eCollection 2015.

Slikker, W., Andersen, M.E., Bogdanffy, M.S., Bus, J.S., Cohen, S.D., Conolly, R.B., David, R.M., Doerrer, N.G., Dorman, D.C., Gaylor, D.W., Hattis, D., Rogers, J.M., Setzer, R.W., Swenberg, J.A., Wallace, K., 2004. Dose-dependent transitions in mechanisms of toxicity. Toxicol. Appl. Pharmacol. 201, 203–225.

Smith, M.T., Guyton, K.Z., Gibbons, C.F., Fritz, J.M., Portier, C.J., Rusyn, I., et al., 2016. Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. Environ. Health Perspect. 124, 713–721.

Sonich-Mullin, C., Fielder, R., Wiltse, J., Baetcke, K., Dempsey, J., Fenner-Crisp, P., Grant, D., Hartley, M., Knaap, A., Kroese, D., Mangelsdorf, I., Meek, E., Rice, J.M., Younes, M., 2001. IPCS conceptual framework for evaluating a mode of action for chemical carcinogenesis. Regul. Toxicol. Pharmacol. 34, 146–152.

Sokal, R.R., Rohlf, F.J., 1981. Biometry: the Principles and Practices of Statistics in Biological Research, second ed. ISBN: 9780716712541.

Thomas, R.S., Black, M., Li, L., Healy, E., Chu, T.-M., Bao, W., et al., 2012. A comprehensive statistical analysis of predicting in vivo hazard using high-throughput *in vitro* screening. Toxicol. Sci. 128, 398–417.

United States Environmental Protection Agency (USEPA), 2005. Guidelines for carcinogen risk assessment. https://www.epa.gov/sites/production/files/2013-09/documents/cancer_guidelines_final_3-25-05.pdf. (Accessed 24 August 2017).

United States Environmental Protection Agency (USEPA), 2015. Systematic Review Relating to Mechanistic Data: what Is Really Needed, and How Can it Be Efficiently Applied? http://ofmpub.epa.gov/eims/eimscomm.getfile?p_download_id=526748. (Accessed 24 August 2017).

United States Environmental Protection Agency (USEPA), 2016. Annual Cancer Report. Chemicals Evaluated for Carcinogenic Potential, Office of Pesticide Programs. http://npic.orst.edu/chemicals_evaluated.pdf. (Accessed 24 August 2017).

United States Environmental Protection Agency (USEPA), 2017. External Review Draft, Toxicological Review of Ethyl Tertiary Butyl Ether. June 2017. http://ofmpub.epa.gov/eims/eimscomm.getfile?p_download_id=531514. (Accessed 24 August 2017).

van Bilsen, J.H.M., Sienkiewicz-Szlapka, E., Lozano-Ojalvo, D., et al., 2017. Application of the adverse outcome pathway (AOP) concept to structure the available in vivo and in vitro mechanistic data for allergic sensitization to food proteins. Clin. Transl. Allergy 7 (13), 1–18.

Zhang, Q., Bhattacharya, S., Conolly, R.B., Clewell, H.J., Kaminski, N.E., Andersen, M.E., 2014. Molecular signaling network motifs provide a mechanistic basis for cellular threshold responses. Environ. Health Perspect. 122, 1261–1270.